

Bayesian Effect Fusion for Categorical Predictors

Daniela Pauger, Helga Wagner

March 31, 2017

Abstract

In this paper, we propose a Bayesian approach to obtain a sparse representation of the effect of a categorical predictor in regression type models. As the effect of a categorical predictor is captured by a group of level effects, sparsity cannot only be achieved by excluding single irrelevant level effects but also by excluding the whole group of effects associated to a predictor or by fusing levels which have essentially the same effect on the response. To achieve this goal, we propose a prior which allows for almost perfect as well as almost zero dependence between level effects a priori. We show how this prior can be obtained by specifying spike and slab prior distributions on all effect differences associated to one categorical predictor and how restricted fusion can be implemented. An efficient MCMC method for posterior computation is developed. The performance of the proposed method is investigated on simulated data. Finally, we illustrate its application on real data from EU-SILC.

keywords: spike and slab prior, sparsity, nominal and ordinal predictor, regression model, MCMC, Gibbs sampler

1 Introduction

In many applications, especially in medical, social or economic studies, potential covariates collected for a regression analysis are categorical, measured either on an ordinal or on a nominal scale. The usual strategy for modelling the effect of categorical covariates is to define one level as baseline and to use dummy variables for the effects of the other levels with respect to this baseline. Hence, the effect of a categorical covariate is captured not by a single but by a group of regression effects. Including categorical variables as covariates in regression type models can therefore easily lead to a high-dimensional vector of regression effects. Moreover, since only the subset of observations with a specific level contribute information to estimation of its effect, estimated effects of rare levels will be associated with high uncertainty.

Many methods have been proposed to achieve sparser models by identifying regressors with non-zero effects. Whereas frequentist methods, e.g. the lasso (Tibshirani, 1996) or the elastic net (Zou and Hastie, 2005) rely on penalties, Bayesian variable selection methods are based on specification of appropriate prior distributions, e.g. shrinkage priors (Park and Casella, 2008; Griffin and Brown, 2010) or spike and slab priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1997; Ishwaran et al., 2001). However, variable selection methods perform selection of single regression effects and are not appropriate for a categorical covariate with more categories, as the natural grouping of the dummy variables capturing its effect is not taken into account.

Moreover, a sparser representation of the effect of a categorical covariate can be achieved not only by restricting all of its level effects to zero but also when some of the levels have the same effect. In this paper, we propose a Bayesian approach to achieve a sparser representation of the effects of a categorical predictor, which encourages both shrinkage of non-relevant effects to zero as well as fusion of (almost) identical level effects.

Approaches that explicitly address inclusion or exclusion of groups of regression coefficients associated to one variable are the group lasso (Yuan and Lin, 2006) and the Bayesian group lasso (Raman et al., 2009; Kyung et al., 2010). Chipman (1996) uses spike and slab priors for grouped selection of the set of all dummy variables related to a categorical predictor. Whereas all these methods aim at sparsity for groups of regression coefficients, the recently proposed sparse-group lasso (Simon et al., 2013) addresses also sparsity within groups by shrinking negligible effects to zero, however not by fusing identical level effects.

To encourage both sparsity of regression effects as well as their differences in regression models with metric predictors, Tibshirani et al. (2005) proposed the fused lasso and Kyung et al. (2010) its Bayesian counterpart, the Bayesian fused lasso. Both methods assume some ordering of effects and shrink only effect differences of subsequent effects to zero. Hence, they are not appropriate for nominal predictors where any effect difference should be subject to shrinkage. Effect fusion for nominal predictors is considered only in Bondell and Reich (2009) who propose a modification of the fused lasso for ANOVA, by Gertheiss and Tutz (Gertheiss and Tutz, 2009; Gertheiss et al., 2011; Gertheiss and Tutz, 2010; Tutz and Gertheiss, 2016) who specify different lasso-type penalties for ordinal and nominal covariates and recently in Tutz and Berger (2014) where tree-structured clustering of effects of categorical covariates is performed.

We address the problem of sparsity for effects of categorical predictors from a Bayesian point of view and incorporate structure in the prior on the regression effects. As the goal is to learn whether two level effects are almost equal or considerably different, we do not specify a standard independence Normal prior for the level effects but explicitly model dependence in their joint precision matrix by allowing for either almost perfect or low dependence. We show that the prior can alternatively be achieved by specifying spike and slab prior distributions on all level effects and their differences and taking into account their linear dependence. Spike and slab prior distributions have been applied extensively in Bayesian approaches to variable selection. The mixture structure with a spike at zero and a flat slab allows for intrinsic classification of effects and effect differences as (almost) zero, when a coefficient is assigned to the spike and as non-zero otherwise. Whereas for a categorical predictor any two level effects will be subject to fusion, it seems natural to exploit the ordering information available for an ordinal predictor by restricting fusion to adjacent categories, which is easily accomplished in our framework. Generally, our proposed method is not limited to categorical predictors but can be applied to all groups of covariates

The rest of the paper is organised as follows: in Section 2 we introduce the data model and in Section 3 the prior distribution constructed to encourage a sparse representation of covariate effects. Posterior inference is discussed in Section 4 and Section 5 investigates the performance of the method for simulated data. An applications of the proposed method for Bayesian effect fusion is illustrated on a real data example in Section 6 and we conclude with Section 7.

2 Model specification

We consider a standard linear regression model with Normal response y and p categorical covariates C_h , $h = 1, \dots, p$, where covariate C_h has $c_h + 1$ ordered or unordered levels $0, \dots, c_h$. To represent its effect on the response y , we define $C_h = 0$ as the baseline category and introduce dummy variables, $X_{h,k}$, to capture the effect of level $C_h = k$ with respect to the baseline category. The regression model is then given as

$$y = \mu + \sum_{h=1}^p \sum_{k=1}^{c_h} X_{h,k} \beta_{h,k} + \varepsilon, \quad (1)$$

where μ is the intercept, $\beta_{h,k}$ is the effect of level k of covariate C_h (with respect to the reference category) and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the error term.

For an $(n \times 1)$ response vector $\mathbf{y} = (y_1, \dots, y_n)'$ we write the model as

$$\mathbf{y} = \mathbf{1}\mu + \sum_{h=1}^p \mathbf{X}_h \boldsymbol{\beta}_h + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (2)$$

where \mathbf{X}_h is the $(n \times c_h)$ design matrix for covariate C_h , $\boldsymbol{\beta}_h$ is the $(c_h \times 1)$ vector of the corresponding regression effects and $\boldsymbol{\varepsilon}$ the $(n \times 1)$ vector of error terms. $\mathbf{1}$ denotes a vector with elements 1 and \mathbf{I} the identity matrix.

3 Prior specification

Bayesian model specification is completed by assigning prior distributions to all model parameters. We assume a prior of the structure

$$p(\mu, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p, \sigma^2) = p(\sigma^2) p(\mu | \sigma^2) \prod_{h=1}^p p(\boldsymbol{\beta}_h | \boldsymbol{\xi}_h) p(\boldsymbol{\xi}_h),$$

where $\boldsymbol{\xi}_h = (\tau_h^2, \boldsymbol{\delta}_h)$ denotes additional hyperparameters, which are specified below. We assign a flat proper prior $p(\mu) \sim \mathcal{N}(0, M_0)$ to the intercept and an Inverse Gamma distribution $p(\sigma^2) \sim \mathcal{G}^{-1}(s_0, S_0)$ to the error variance.

The prior on the regression effects $\boldsymbol{\beta}_h$ is specified hierarchically as

$$\boldsymbol{\beta}_h \sim \mathcal{N}(\mathbf{0}, \gamma_h \tau_h^2 \mathbf{Q}_h^{-1}(\boldsymbol{\delta}_h)) \quad (3)$$

$$\tau_h^2 \sim \mathcal{G}^{-1}(g_{h0}, G_{h0}), \quad (4)$$

where γ_h is a fixed constant, τ_h^2 is a scale parameter and the matrix \mathbf{Q}_h determines the structure of the prior precision matrix. To encourage effect fusion, we let \mathbf{Q}_h depend on a vector $\boldsymbol{\delta}_h$ of indicator variables $\delta_{h,kj}$, which are defined for each pair of level effects k and j subject to fusion. $\delta_{h,kj} = 1$ indicates that $\beta_{h,k}$ and $\beta_{h,j}$ differ considerable and hence two regression parameters are needed to capture their respective effects whereas for $\delta_{h,kj} = 0$ the effects are almost identical and the two level effects could be fused. To allow fusion of level effects to 0, i.e. conventional variable selection, we include in $\boldsymbol{\delta}_h$ also indicators $\delta_{h,k0}$, $k = 1, \dots, c_h$.

The dimension of $\boldsymbol{\delta}_h$ and the concrete specification of \mathbf{Q}_h depend on which pairs of effects are subject to fusion. We discuss the case where fusion is completely unrestricted and hence any pair of effects might be fused in Section 3.1. Whereas unrestricted effect fusion will be appropriate for a nominal covariate, for an ordinal covariate information on the ordering of levels is available which suggests to fuse only adjacent categories as discussed in Gertheiss and Tutz (2009). We describe effect fusion taking into account restrictions that preclude direct fusion for specified pairs of effects in Section 3.2. For notational convenience we define $\beta_{h,0} = 0$ and drop the covariate index h in the following.

3.1 Prior for unrestricted effect fusion

For unrestricted effect fusion, we introduce an indicator δ_{kj} for each pair of effects $k = 1, \dots, c$ and $j = 0, \dots, k-1$ (including 0 for the baseline) and hence $\boldsymbol{\delta}$ is of dimension $d = \binom{c+1}{2}$. We define

$$\kappa_{kj} = \delta_{kj} + r(1 - \delta_{kj}),$$

where r is a fixed large number (e.g. $r = 10000$) for $k > j$ and $\kappa_{jk} = \kappa_{kj}$ for $j > k$.

The structure of the prior precision matrix is then specified as

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} \sum_{j \neq 1} \kappa_{1j} & -\kappa_{12} & \dots & -\kappa_{1c} \\ -\kappa_{21} & \sum_{j \neq 2} \kappa_{2j} & \dots & -\kappa_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ -\kappa_{c1} & -\kappa_{c2} & \dots & \sum_{j \neq c} \kappa_{cj} \end{pmatrix}. \quad (5)$$

and finally, we set $\gamma = c/2$.

The structure matrix $\mathbf{Q}(\boldsymbol{\delta})$ determines the prior precision matrix of $\boldsymbol{\beta}$ up to the scale factor $\gamma\tau^2$ and therefore has to be symmetric and positive definite. Symmetry of $\mathbf{Q}(\boldsymbol{\delta})$ is guaranteed by definition and positive definiteness as

$$\boldsymbol{\beta}'\mathbf{Q}(\boldsymbol{\delta})\boldsymbol{\beta} = \sum_{k=1}^c \beta_k^2 \kappa(\delta_{k0}) + \sum_{k=1}^c \sum_{j < k} (\beta_k - \beta_j)^2 \kappa(\delta_{kj}) > 0, \quad (6)$$

if $\boldsymbol{\beta} \neq \mathbf{0}$, see Appendix A.1 for a detailed proof.

To interpret the structure matrix $\mathbf{Q}(\boldsymbol{\delta})$, we subsume in $\boldsymbol{\delta}_k = (\delta_{k0}, \dots, \delta_{k,k-1}, \delta_{k+1,k}, \dots, \delta_{c,k})$ all indicators related to level k . The diagonal elements q_{kk} of $\mathbf{Q}(\boldsymbol{\delta})$ determine the partial prior precisions and the off-diagonal elements q_{kj} the partial prior correlations of the level effects:

$$\text{Cor}(\beta_k, \beta_j | \boldsymbol{\beta}_{\setminus kj}) = -\frac{q_{kj}}{\sqrt{q_{jj}q_{kk}}} \quad (7)$$

$$\text{Prec}(\beta_k | \boldsymbol{\beta}_{\setminus k}) = q_{kk}/(\gamma\tau^2). \quad (8)$$

Thus, the prior allows for either high (if $\delta_{kj} = 0$) or low (if $\delta_{kj} = 1$) positive prior partial correlation. Further, depending on $\boldsymbol{\delta}_k$, c different values of the prior precision are possible, ranging from $\frac{c}{\gamma\tau^2}$ (for $\boldsymbol{\delta}_k = \mathbf{1}$) to $\frac{cr}{\gamma\tau^2}$ (for $\boldsymbol{\delta}_k = \mathbf{0}$). As an example, consider a covariate with $c = 3$ levels, where $\delta_{kj} = 1$ for all $k > j = 0, \dots, c$ except one and $r = 10000$. If $\delta_{10} = 0$ the structure matrix is

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} 10002 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}$$

and the marginal prior on β_1 is concentrated close to zero. For $\delta_{32} = 0$,

$$\mathbf{Q}(\boldsymbol{\delta}) = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 10002 & -10000 \\ -1 & -10000 & 10002 \end{pmatrix}$$

and hence the joint prior on (β_2, β_3) is concentrated close to $\beta_2 = \beta_3$. Note that marginally the prior on off-diagonal elements of \mathbf{Q} is a mixture of two inverse Gamma distributions.

The structure of the quadratic form in equation (6) suggests a straightforward interpretation of the effect fusion prior in terms of Normal priors on all effect differences $\theta_{kj} = \beta_k - \beta_j$, $k = 1, \dots, c$; $j = 0, \dots, k-1$. For $\delta_{kj} = 0$, the effect difference θ_{kj} is concentrated around zero, whereas it is more dispersed for $\delta_{kj} = 1$. Actually, as we show in Appendix A.2 the effect fusion prior

specified above can be derived by starting from independent spike and slab priors on all effect differences θ_{kj} and then correcting for the linear restrictions $\theta_{kj} = \theta_{k0} - \theta_{j0}$.

Finally, we note that from a frequentist perspective, the effect fusion prior can be interpreted as an adaptive quadratic penalty (see equation (6)), with either heavy or slight penalization of effect differences. In contrast, Gertheiss and Tutz (2010) use a weighted L_1 penalty on the effect differences.

3.2 Prior for restricted effect fusion

If information on the structure of the level effects is available, this information can be exploited by allowing only fusion of specific pairs of level effects. Consider e.g. an ordinal covariate where the ordering of levels suggests to allow only fusion of subsequent level effects β_{k-1} and β_k i.e. restricting direct fusion of effects to adjacent categories. A restriction that e.g. β_k and β_j should not be fused can be implemented in our prior in two ways: we can either fix the indicator $\delta_{kj} = 0$ or directly set the corresponding element in the prior precision matrix $\mathbf{Q}(\boldsymbol{\delta})$, $q_{kj} = 0$. Setting $\delta_{kj} = 0$ implies that effects β_k and β_j are still smoothed to each other and hence a soft restriction, whereas $q_{kj} = 0$ is a hard restriction which implies conditional independence of β_k and β_j .

Whereas implementation of soft restrictions is straightforward, (hard) conditional independence restrictions require slight modifications for the structure matrix $\mathbf{Q}(\boldsymbol{\delta})$, the vector of indicators $\boldsymbol{\delta}$ and the constant γ . We start by introducing a vector $\boldsymbol{\zeta}$ of indicators ζ_{kj} , which are defined for each effect difference θ_{kj} . The elements of $\boldsymbol{\zeta}$ are fixed and indicate whether an effect difference is subject to fusion (for $\zeta_{kj} = 1$) or not (for $\zeta_{kj} = 0$). Deviating from unrestricted effect fusion considered in Section 3.1, we define a stochastic indicator δ_{kj} only for those effect differences where $\zeta_{kj} = 1$ and hence the dimension of $\boldsymbol{\delta}$ is $d = \sum_{k=1}^c \sum_{0 \leq j < k} \zeta_{kj}$.

To allow off-diagonal elements of the prior precision to be zero, we set

$$q_{kj} = \begin{cases} -\kappa_{kj} & \text{if } \zeta_{kj} = 1 \\ 0 & \text{if } \zeta_{kj} = 0 \end{cases}$$

and $q_{jk} = q_{kj}$. Thus q_{kj} takes the value zero if $\zeta_{kj} = 0$ and $-\kappa_{kj}$ otherwise.

Similarly, the diagonal elements are specified as

$$q_{kk} = \begin{cases} \kappa_{k0} - \sum_{k \neq j} q_{kj} & \text{if } \zeta_{k0} = 1 \\ - \sum_{k \neq j} q_{kj} & \text{if } \zeta_{k0} = 0. \end{cases}$$

As noted above, an important special case is an ordinal covariate where it is natural to restrict fusion to adjacent categories as in Gertheiss and Tutz (2009), i.e.

$$\zeta_{kj} = \begin{cases} 1 & \text{for } j = k - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the vector of indicators $\boldsymbol{\delta}$ has only $d = c$ elements and $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ is a tri-diagonal matrix with elements

$$\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta}) = \begin{pmatrix} \kappa_{10} + \kappa_{21} & -\kappa_{21} & 0 & \dots & 0 \\ -\kappa_{21} & \kappa_{21} + \kappa_{32} & -\kappa_{32} & \dots & 0 \\ 0 & -\kappa_{32} & \kappa_{32} + \kappa_{43} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & \kappa_{c,c-1} \end{pmatrix}. \quad (9)$$

In this case, the maximum value of a diagonal element q_{jj} is two and therefore we set $\gamma = 1$.

It is easy to show that this specification of $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ corresponds to a random walk prior with initial value $\beta_0 = 0$ on the regression effects:

$$\beta_k = \beta_{k-1} + \theta_k, \quad \theta_k \sim \mathcal{N}(0, \tau^2 / \kappa_{k,k-1}).$$

Due to the spike and slab structure, this prior allows for adaptive smoothing, with almost no smoothing for $\delta_{k,k-1} = 1$ and pronounced smoothing for $\delta_{k,k-1} = 0$.

Another special case is the standard spike and slab prior used for variable selection, where only fusion of level effects to the baseline, i.e. shrinkage of β_k to zero is considered. The spike and slab prior is recovered in our framework when $\gamma = 1$,

$$\zeta_{kj} = \begin{cases} 1 & \text{for } j = 0 \\ 0 & \text{otherwise} \end{cases}$$

and hence non-diagonal elements of $\mathbf{Q}(\boldsymbol{\zeta}, \boldsymbol{\delta})$ are zero and $q_{kk} = \kappa_{k0}$.

3.3 Prior on the indicator variables

A standard choice in variable selection is to assume conditional prior independence of the elements of $\boldsymbol{\delta}$ with $p(\delta_{kj} = 1) = \omega$, where ω is either fixed or assigned a hyperprior $\omega \sim \mathcal{B}(v_0, w_0)$.

This would in principle be possible also with our prior, however, from a computational point of view a more convenient choice is to set

$$p(\boldsymbol{\delta}) \propto |\mathbf{Q}(\boldsymbol{\delta})|^{-1/2} r^{\sum(1-\delta_{kj})/2}$$

as with this choice the determinant of $\mathbf{Q}(\boldsymbol{\delta})$ cancels out in the joint prior $p(\boldsymbol{\beta}, \boldsymbol{\delta} | \tau^2)$, which results as

$$p(\boldsymbol{\beta}, \boldsymbol{\delta} | \tau^2) = p(\boldsymbol{\beta} | \boldsymbol{\delta}, \tau^2) p(\boldsymbol{\delta}) = \left(\frac{1}{\gamma \tau^2}\right)^{c/2} \exp\left(-\frac{\boldsymbol{\beta}' \mathbf{Q}(\boldsymbol{\delta}) \boldsymbol{\beta}}{2\gamma \tau^2}\right) (\sqrt{r})^{\sum(1-\delta_{kj})}. \quad (10)$$

3.4 Choice of hyperparameters

The hyperparameters of the effect fusion prior should be chosen to minimize the expected loss of the underlying decision problem: loss occurs if level effects which are different are fused or effects which are equal are not fused. We call the first case *false negative*, as a non-zero effect difference is not detected and the second *false positive*, as a zero effect difference is classified as non-zero. False positives and false negatives have different impacts: if an effect difference is falsely classified as positive, two parameters are included in the model though only one would be sufficient. This results in a loss of estimation efficiency. In contrast, if an effect difference is falsely classified as negative, two effects that are actually different from each other are modelled by only one parameter. This will result in biased estimation and bad prediction performance. Hence, the primary goal will be to avoid false negatives.

From the representation of the prior in terms of spike and slab priors on all effect differences θ_{kj} , it is evident that the conditional prior fusion probability $P(\delta_{kj} = 0 | \theta_{kj})$ depends on the slab to spike ratio r and the parameters of the inverse Gamma distribution for the slab variance g_0 and G_0 .

We propose to set $g_0 = 5$, a standard choice in variable selection (see e.g. Fahrmeir et al. (2010); Scheipl et al. (2012)), where the tails of spike and slab are not too thin to cause mixing problems in MCMC. For fixed $\theta_{kj} > 0$, the prior fusion probability $P(\delta_{kj} = 0 | \theta_{kj})$ is lower for a larger slab

to spike ratio r and for a smaller scale parameter G_0 . This suggests to choose a high value for r and a small value for G_0 . However, shrinkage of effect differences to zero is more pronounced with a smaller scale parameter G_0 , also under the slab, which might hamper detection of small effect differences. We will investigate this issue in more detail in the simulation study in Section 5.3.

4 Posterior inference

4.1 MCMC scheme

We subsume the regression coefficients in $\boldsymbol{\beta} = (\mu, \beta_1, \dots, \beta_p)$ and denote the collection of all parameters by $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\tau}^2, \sigma^2)$. The goal is posterior inference for $\boldsymbol{\Theta}$, which can be accomplished by sampling from the posterior distribution

$$p(\boldsymbol{\Theta}|\mathbf{y}) \propto p(\boldsymbol{\Theta})p(\mathbf{y}|\boldsymbol{\Theta})$$

using MCMC methods.

As the model is a linear Bayesian regression model with a conditionally conjugate prior, MCMC is straightforward. After choosing starting values for $\boldsymbol{\xi} = (\boldsymbol{\tau}^2, \boldsymbol{\delta})$ and σ^2 MCMC proceeds by iterating between the following steps:

1. Update the prior variance matrix $\mathbf{B}_0(\boldsymbol{\xi})$ and sample the regression coefficients $\boldsymbol{\beta}$ from the full conditional Normal distribution $\mathcal{N}(\mathbf{b}, \mathbf{B})$ with moments given as

$$\begin{aligned}\mathbf{B}^{-1} &= \frac{1}{\sigma^2} \mathbf{X}'\mathbf{X} + (\mathbf{B}_0(\boldsymbol{\xi}))^{-1} \\ \mathbf{b} &= \frac{1}{\sigma^2} \mathbf{B}\mathbf{X}'\mathbf{y},\end{aligned}$$

where $\mathbf{B}_0(\boldsymbol{\xi})$ is block-diagonal with first element M_0 (for the intercept) and the matrices $\mathbf{B}_{0h}(\boldsymbol{\xi}_h) = \tau_h^2 \gamma_h \mathbf{Q}(\boldsymbol{\delta}_h)^{-1}$.

2. Sample the error variance σ^2 from the Inverse Gamma distribution $\mathcal{G}^{-1}(s, S)$ with parameters

$$\begin{aligned}s &= s_0 + n/2 \\ S &= S_0 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

3. For $h = 1, \dots, p$: sample the scale parameter τ_h^2 from the Inverse Gamma distribution $\mathcal{G}^{-1}(g_h, G_h)$ with parameters

$$g_h = g_{h0} + \frac{c_h}{2}$$

$$G_h = G_{h0} + \frac{1}{2\gamma_h} \beta_h' \mathbf{Q}(\boldsymbol{\delta}_h) \beta_h.$$

4. As both the likelihood given in equation (6) as well as the prior $p(\boldsymbol{\delta}_h)$ can be factorized with respect to the indicators $\delta_{h,kj}$, these can be sampled independently from

$$p(\delta_{h,kj} = 1 | \beta_{h,k}, \beta_{h,j}, \tau_h^2) = \frac{1}{1 + \frac{p(\delta_{h,kj}=0 | \beta_{h,k}, \beta_{h,j}, \tau_h^2)}{p(\delta_{h,kj}=1 | \beta_{h,k}, \beta_{h,j}, \tau_h^2)}} = \frac{1}{1 + L_{h,kj}},$$

where

$$L_{h,kj} = \sqrt{r} \exp \left(- \frac{r-1}{2} \frac{(\beta_{h,k} - \beta_{h,j})^2}{\gamma_h \tau_h^2} \right).$$

4.2 Model selection

Effect fusion aims at selecting an appropriate model for a categorical predictor and thus is a particular model selection problem. In a Bayesian approach, model selection is usually based on posterior model probabilities and the goal is to find the model with maximum posterior model probability. Slightly differing, in Bayesian variable selection typically not the maximum probability model but the median probability model, i.e. the model including all covariates that have an estimated posterior inclusion probability larger than 0.5, is selected.

To select a model with potentially fused effects one could use the estimated fusion probabilities $\hat{\psi}_{h,kj} = 1 - \hat{\delta}_{h,kj}$ where $\hat{\delta}_{h,kj}$ is the mean of the corresponding MCMC draws, and fuse effects if $\hat{\psi}_{h,kj} > 0.5$. However, this strategy could yield a logically inconsistent model where e.g. levels j and l as well as k and l are fused but not levels j and k . Hence, we fuse levels k and j only if $\hat{\psi}_{h,kj} > 0.5$ and for all $l \neq j, k$ both $\hat{\psi}_{h,kl}$ and $\hat{\psi}_{h,jl}$ are either larger or smaller than 0.5. This strategy avoids logically inconsistent models and as levels are only fused when evidence is clear, takes into account the asymmetry in loss of false positives and false negatives.

After model selection, we estimate the dummy-coded regression coefficients of the selected model by a Bayesian regression under a flat Normal prior $\mathcal{N}(0, \mathbf{I}B_0)$ with $B_0 = 10000$ on all effects.

5 Simulation study

We now illustrate the performance of the proposed method in a simulation study and compare our method to various other approaches: the regularization approach in Gertheiss and Tutz (2010) (*Penalty*), the Bayesian lasso (*BLasso*), Bayesian elastic net (*BEN*) and the group lasso (*GLasso*). Additionally, we include Bayesian regularization via graph Laplacian (*GLap*), proposed in Liu et al. (2014). They also specify the prior directly on the elements of the prior precision matrix, with the goal to identify conditional independence by shrinking off-diagonal elements to zero.

A list of the used R packages and related papers are given in the Appendix B.1. Additionally, we fit the full model 1 (*Full*) with separate dummy variables for each level and the true model (*True*), i.e. the model with fused categories according to data generation. We use a set-up similar as in Gertheiss and Tutz (2010) and compare the methods with respect to parameter estimation, predictive performance and model selection.

5.1 Simulation set-up

For the simulation study we generate 100 data sets with $n = 500$ observations from the Gaussian linear regression model (2) with intercept $\mu = 1$, error variance $\varepsilon \sim \mathcal{N}(0, 1)$ and fixed design matrix \mathbf{X} . We use four ordinal and four nominal predictors, where two regressors have eight and two have four categories for each type of covariate (ordinal and nominal). Regression effects are set to $\beta_1 = (0, 1, 1, 2, 2, 4, 4)$ and $\beta_3 = (0, -2, -2)$ for the ordinal and $\beta_5 = (0, 1, 1, 1, 1, -2, -2)$ and $\beta_7 = (0, 2, 2)$ for the nominal covariates, and $\beta_h = \mathbf{0}$ for $h = 2, 4, 6, 8$. Levels of the predictors are generated with probabilities $\pi = (0.1, 0.1, 0.2, 0.05, 0.2, 0.1, 0.2, 0.05)$ and $\pi = (0.1, 0.4, 0.2, 0.3)$ for regressors with eight and four levels, respectively.

To perform effect fusion, we specify a Normal prior with variance $B_0 = 10000$ for the intercept and the improper prior $p(\sigma^2) \propto 1/\sigma^2$ (which corresponds to an Inverse Gamma distribution with parameters $s_0 = S_0 = 0$) for the error variance σ^2 . For each covariate C_h , the hyperparameters are set to $G_{h0} = 20$ and $r = 20000$, but we investigate also different values in Section 5.3.

MCMC is run for 10000 iterations after burnin of 5000 to perform model selection for each data set. Models *Full* and *True* and the refit of the selected model are estimated under a flat Normal prior $\mathcal{N}(0, \mathbf{I}B_0)$ with $B_0 = 10000$ on the regression coefficients and MCMC is run for 3000 iterations after a burnin

of 1000. The tuning parameters of the frequentist methods *Penalty* and *GLasso* are selected automatically via cross-validation in the corresponding R packages. For the Bayesian methods, we use the default prior parameter settings in the code (for *GLap*) and the R packages *monomvn* and *EBglmNet* and estimate the regression coefficients by the posterior means.

5.2 Simulation results

We first compare the suggested method for Bayesian effect fusion to the other approaches with respect to estimation of the regression effects. Figure 1 shows the mean squared estimation error (MSE) defined for each covariate C_h as

$$MSE_h^{(i)} = \frac{1}{c_h} \sum_{k=1}^{c_h} (\hat{\beta}_{h,k}^{(i)} - \beta_{h,k})^2.$$

Obviously, the mean of the MSEs (over all 100 data sets) are lower for Bayesian effect fusion than for all other methods. Bayesian effect fusion performs particularly well for covariates where all levels have an effect of zero (variables 2, 4, 6, 8). For covariates with non-zero effects overall performance is good, but for some data sets the MSE can be higher than for the full model, when levels with actually different effects are fused, see eg. covariate 5, a nominal covariate with eight levels.

The competitors *BLasso* and *BEN* perform very well both for covariates with zero as well as covariates with non-zero effects. *Penalty* which is designed for effect fusion does not clearly outperform these two methods for covariates with non-zero effects but yields higher MSEs for covariates with no effects. *GLasso* performs reasonably well for covariates with no effects but worse for covariates with non-zero effects and *GLap* yields only slight improvements compared to the full model for covariates with no effect.

We would like to remark that also model averaged estimates, which are obtained as posterior mean estimates from the first MCMC run under the effect fusion prior, perform very well with respect to parameter estimation.

To evaluate the predictive performance of Bayesian effect fusion, we generate a new sample of $n^* = 500$ observations z_j , $j = 1, \dots, n^*$ from the linear regression model (2) with fixed regressors $\tilde{\mathbf{x}}_j$ and the same parameters as in the simulated data sets. Predictions for these new observations are computed using the estimates from each of the original data sets as $\hat{z}_j^{(i)} = \tilde{\mathbf{x}}_j \hat{\boldsymbol{\beta}}^{(i)}$, $i = 1, \dots, 100$.

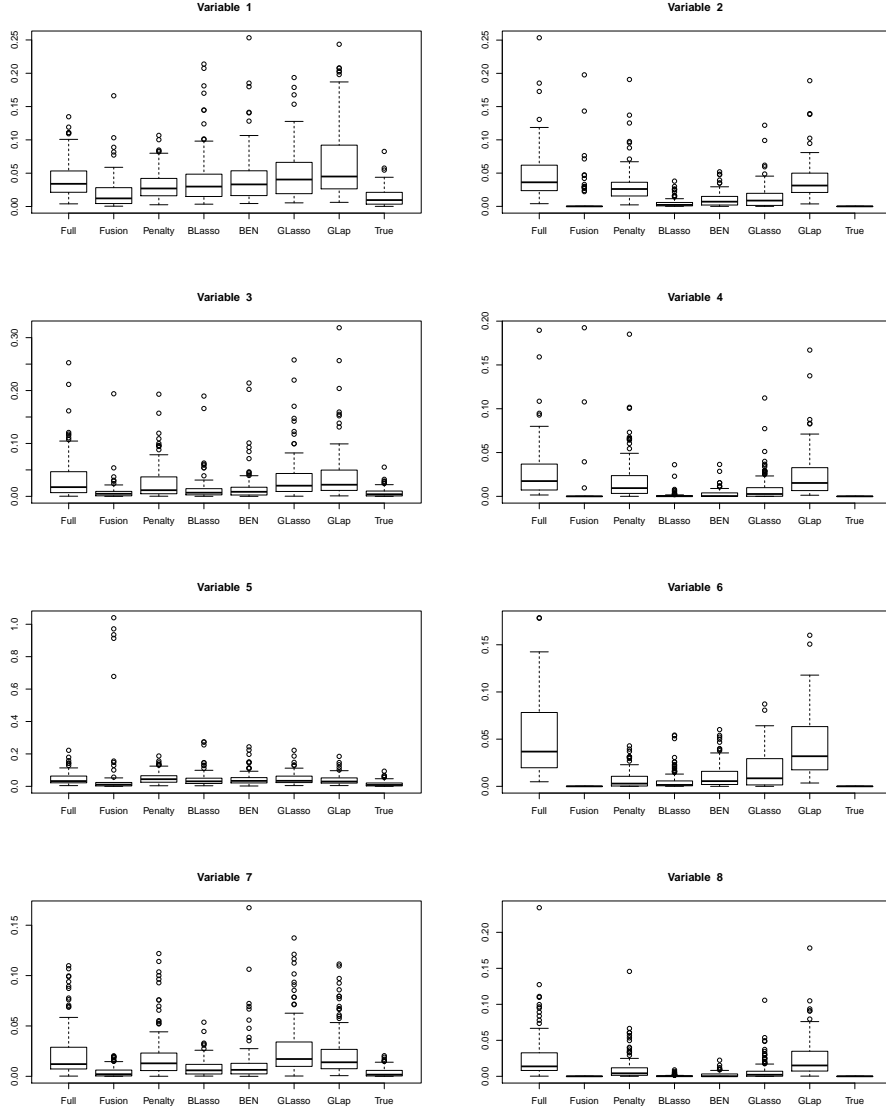


Figure 1: Simulation study: MSEs for 100 simulated data sets. Variable 1 to 4 are ordinal, variable 5 to 8 are nominal. Variables on the right panel (even numbers) have no effect on the response.

The mean squared prediction errors (MSPE) defined for each data set as

$$\text{MSPE}^{(i)} = \frac{1}{n^*} \sum_{j=1}^{n^*} (z_j - \tilde{\mathbf{x}}_j \hat{\boldsymbol{\beta}}^{(i)})^2, \quad i = 1, \dots, 100$$

are shown in Figure 2. The predictive performance for our method is almost as good as if the true model were known and considerably better than for all competing methods in most data sets. *BLasso* is the second best method and also *Penalty*, *GLasso* and *BEN* yield slightly smaller prediction errors compared to full model. Prediction errors using *GLap* are similar to those from the full model.

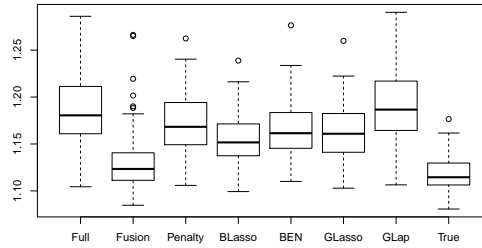


Figure 2: Simulation study: MSPEs of 500 new observations using estimates of 100 simulated data sets.

Finally, to evaluate and compare the performance of the methods with respect to model selection, we use the true positive rate (TPR), the true negative rate (TNR), the positive predictive value (PPV) and the negative predictive value (NPV), see Appendix B.2 for detailed definitions. If fusion is completely correct, all four values are equal to 100% but TPR and PPV are not defined for covariates where all effects are zero. For the effect fusion prior we perform model selection as described in Section 4.2, for the other methods we consider two level effects as identical if the posterior mean of their difference is smaller or equal to 0.01. Results reported in Tables 1 and 2 show that Bayesian effect fusion clearly outperforms all other methods in particular with respect to identifying categories with the same effect (TNR).

5.3 Influence of hyperparameters

In this section, we investigate the sensitivity of model selection under the Bayesian effect fusion prior with respect to the hyperparameters. As discussed in Section 3.4 the primary goal is to avoid incorrect fusion of level

Table 1: Simulated data: Model selection results for ordinal predictors, means of 100 simulated data sets.

Covariate	Method	TPR	TNR	PPV	NPV
1	Fusion	99.7	95.8	95.8	99.8
	Penalty	100	18.8	48.8	100
	BLasso	100	6.8	44.8	100
	BEN	100	19.2	48.5	100
	GLasso	100	2.2	43.5	100
	GLap	100	2.5	43.6	100
2	Fusion	-	97.7	-	100
	Penalty	-	21.3	-	100
	BLasso	-	24.1	-	100
	BEN	-	53.9	-	100
	GLasso	-	21.7	-	100
	GLap	-	2.7	-	100
3	Fusion	100	99.0	99.3	100
	Penalty	100	24.5	44.8	100
	BLasso	100	25.0	43.7	100
	BEN	100	46.5	52.5	100
	GLasso	100	10.5	36.8	100
	GLap	100	6.5	35.5	100
4	Fusion	-	98.7	-	100
	Penalty	-	21.3	-	100
	BLasso	-	48.3	-	100
	BEN	-	61.3	-	100
	GLasso	-	34.0	-	100
	GLap	-	7.3	-	100

effects, i.e. *false negatives* while keeping *false positives* at a moderate level. We therefore focus on false negative rates, $\text{FNR} = 1 - \text{TPR}$ and false positive rates $\text{FPR} = 1 - \text{TNR}$ and report both rates for various values of G_0 and fixed $r = 20000$ in Table 3 and for various values of r and fixed $G_0 = 20$ in Table 4.

Results in Table 3 indicate that increasing G_0 from 0.2 to 200 has little effect on FNR but results in lower FPR for ordinal predictors (covariate 1 to 4), whereas it has little effect on FPR but results in higher FNR for nominal effects. Hence, G_0 should be chosen not too high. From our experience $G_0 = 2$ is a good choice to detect also small effect differences of nominal predictors whereas for ordinal predictors a larger value for G_0 , e.g. $G_0 = 20$ is reasonable.

Table 4 reports FNR and FPR for values of r from $2 \cdot 10^2$ to $2 \cdot 10^5$. Obviously, r has almost no influence for ordinal predictors but for nominal covariates low values of r encourage too much fusion and hence yield a high FNR. These

Table 2: Simulated data: Model selection results for nominal predictors, means of 100 simulated data sets.

Covariate	Method	TPR	TNR	PPV	NPV
5	Fusion	98.3	98.5	99.4	97.0
	Penalty	100	17.2	75.3	100
	BLasso	100	6.8	72.9	100
	BEN	99.9	12.0	74.0	99.5
	GLasso	100	4.2	72.3	100
	GLap	100	4.0	72.3	100
6	Fusion	-	100	-	100
	Penalty	-	51.4	-	100
	BLasso	-	27.5	-	100
	BEN	-	54.5	-	100
	GLasso	-	21.3	-	100
	GLap	-	3.9	-	100
7	Fusion	100	99.5	99.8	100
	Penalty	100	17.5	71.6	100
	BLasso	100	22.5	72.7	100
	BEN	100	43.0	78.3	100
	GLasso	100	5.5	68.1	100
	GLap	100	4.5	67.9	100
8	Fusion	-	100	-	100
	Penalty	-	27.5	-	100
	BLasso	-	47.8	-	100
	BEN	-	70.2	-	100
	GLasso	-	35.3	-	100
	GLap	-	4.7	-	100

results indicate that r should not be too small, but still small enough to avoid stickiness of MCMC. We suggest to use a value of at least $2 \cdot 10^4$.

6 Real data example

As an illustration of Bayesian effect fusion on real data, we model contributions to private retirement pension in Austria. The data are from the European household survey EU-SILC (SILC = Survey on Income and Living Conditions) 2010 in Austria. We use a linear regression model to analyse the effects of socio-demographic variables on the (log-transformed) annual contributions to private retirement pensions. As potential regressors we consider **gender** (binary, 1=female/0=male), **age group** (ordinal with eleven levels), **child in household** (binary, 1=yes/0=no), **income class** (in quartiles of the total data set, i.e. ordinal with four levels), **federal state** of residence

Table 3: Simulated data: Model selection results for $r = 20,000$ and various G_0

Covariate \ G_0	FNR				FPR			
	0.2	2	20	200	0.2	2	20	200
1	0.0	0.3	0.3	0.3	13.5	10.5	4.2	1.7
2	-	-	-	-	12.4	8.7	2.3	0.3
3	0.0	0.0	0.0	0.0	11.5	9.0	1.0	0.5
4	-	-	-	-	17.7	6.3	1.3	0.3
5	0.3	0.3	1.7	78.8	1.2	0.9	1.5	0.0
6	-	-	-	-	0.0	0.6	0.0	0.0
7	0.0	0.0	0.0	0.0	3.5	2.5	0.5	0.0
8	-	-	-	-	0.0	0.0	0.0	0.0

Table 4: Simulated data: Model selection results for $G_0 = 20$ and various r

Covariate \ r	FNR				FPR			
	$2 \cdot 10^2$	$2 \cdot 10^3$	$2 \cdot 10^4$	$2 \cdot 10^5$	$2 \cdot 10^2$	$2 \cdot 10^3$	$2 \cdot 10^4$	$2 \cdot 10^5$
1	0.3	0.3	0.3	0.3	3.2	4.2	4.2	4.2
2	-	-	-	-	0.9	1.6	2.3	2.0
3	0.0	0.0	0.0	0.0	0.5	1.5	1.0	1.0
4	-	-	-	-	0.3	1.0	1.3	1.0
5	100	84.4	1.7	0.5	0.0	0.0	1.5	2.1
6	-	-	-	-	0.0	0.0	0.0	0.6
7	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5
8	-	-	-	-	0.0	0.0	0.0	0.0

in Austria (nominal with nine levels), highest attained level of **education** (nominal with ten levels) and **employment status** (nominal with four levels). We restrict the analysis to observations without missing values in regressors and/or response and a minimum annual contribution of EUR 100. Hence, the final data set used for our analysis comprises 3077 persons.

We standardize the response and fit a regression model including all potential covariates. Results reported in Table 5 indicate that several levels of covariate **education** have a similar effect and most level effects of **federal state** are close to zero, which suggests that a sparser model might be adequate for these data.

To specify the effect fusion prior, we choose the hyperparameters with $r = 50000$, $G_{h0} = 2$ for nominal and $G_{h0} = 20$ for ordinal predictors. To perform model selection, MCMC was run for 50000 iterations after a burn-in of 30000.

Table 5: EU-SILC data: Posterior means and 95% HPD intervals of full model and selected model.

	Full model Posterior mean	95% HPD interval		Selected model Posterior mean	95% HPD interval
Intercept	-1.15	(-1.38 – -0.91)		-1.10	(-1.31 – -0.90)
Age					
20-25	0.20	(-0.03 – 0.42)	}	0.33	(0.14 – 0.52)
25-30	0.36	(0.15 – 0.57)			
30-35	0.60	(0.40 – 0.80)	}	0.65	(0.45 – 0.86)
35-40	0.74	(0.53 – 0.95)			
40-45	0.80	(0.60 – 1.00)	}	0.82	(0.63 – 1.00)
45-50	0.90	(0.70 – 1.10)		0.93	(0.74 – 1.12)
50-55	1.01	(0.80 – 1.23)	}		
55-60	1.06	(0.81 – 1.30)		1.06	(0.86 – 1.25)
60-65	1.23	(0.83 – 1.62)			
> 65	0.67	(0.14 – 1.22)		0.56	(0.09 – 1.05)
Female	-0.24	(-0.32 – -0.17)		-0.25	(-0.31 – -0.18)
Child	0.00	(-0.07 – 0.07)		-	-
Income					
1st quartile	0.20	(0.07 – 0.32)	}		
2nd quartile	0.25	(0.13 – 0.37)		0.23	(0.12 – 0.34)
3rd quartile	0.52	(0.40 – 0.64)		0.53	(0.42 – 0.65)
Federal State					
Carinthia	-0.16	(-0.30 – -0.01)		-	-
Lower Austria	0.06	(-0.03 – 0.16)		-	-
Burgenland	-0.03	(-0.21 – 0.15)		-	-
Salzburg	0.15	(0.01 – 0.30)		-	-
Styria	0.01	(-0.10 – 0.13)		-	-
Tyrol	0.08	(-0.06 – 0.20)		-	-
Vorarlberg	0.02	(-0.15 – 0.19)		-	-
Vienna	0.00	(-0.11 – 0.10)		-	-
Education					
Apprenticeship, trainee	0.09	(-0.04 – 0.23)	}	0.00	-
Master craftman's diploma	0.24	(0.06 – 0.43)			
Nurse's training school	0.22	(-0.04 – 0.47)			
Other vocational school (medium level)	0.26	(0.11 – 0.42)			
Academic secondary school (upper level)	0.22	(0.05 – 0.37)		0.21	(0.14 – 0.28)
College for higher vocational education	0.28	(0.12 – 0.43)			
Vocational school for apprentices	0.29	(0.06 – 0.51)			
University, academy: first degree	0.35	(0.20 – 0.49)			
University: doctoral studies	1.12	(0.85 – 1.37)		1.05	(0.82 – 1.27)
Employment status					
Unemployed	-0.10	(-0.34 – 0.12)		-	-
Retired	-0.15	(-0.37 – 0.07)		-	-
Not-working (other reason)	0.01	(-0.11 – 0.14)		-	-

Figure 3 shows the estimated posterior means of the pairwise fusion probabilities $\hat{\psi}_{h,kk-1}$ for the ordinal covariate **age group**. The estimated fusion probability is higher than 0.5 (dotted line) for four levels, which indicates that age categories 25 – 30 and 30 – 35 could be fused to a category 25 – 35. Similarly 35 – 40 and 40 – 45 could be fused to 35 – 45 and the three categories

50 – 55, 55 – 60 and 60 – 65 to a new category 50 – 65.

The estimated fusion probabilities for the nominal covariate **education** are displayed in Figure 4. All pairwise fusion probabilities, including those for fusion to baseline category 0 are given in form of a symmetric matrix, where darker colours indicate higher fusion probabilities. The values in the diagonal represent fusion probability of a category with itself and hence are always one. Obviously, only three levels are required to capture the effect of education (secondary school and apprenticeship; doctoral degree and all remaining levels) reducing the number of effects from nine to two.

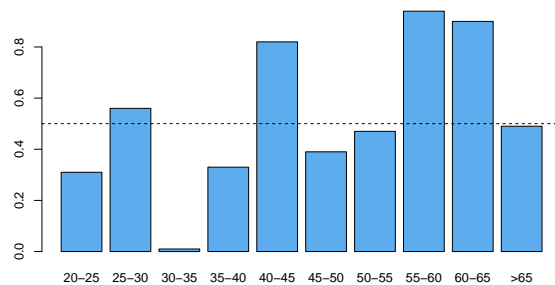


Figure 3: Covariate **age group**: estimated probabilities for fusion with preceding level

Based on the estimated pairwise fusion probabilities we perform model selection for all covariates as described in Section 4.2. Covariates **child**, **federal state** and **employment status** are completely excluded from the model and levels of the covariates **age group**, **income class** and **education** are fused. Thus, the selected model has only eleven regression effects compared to 35 in the full model. Results of a refit of the selected model using flat priors are shown in the right panel of Table 5. The posterior mean of the error variance, $\hat{\sigma}^2 = 0.828$, is almost identical to that of the full model, where $\hat{\sigma}^2 = 0.826$.

7 Conclusion

In this paper, we present a method for sparse modelling of the effects of categorical covariates in Bayesian regression models. Sparsity is achieved by excluding irrelevant predictors and/or by fusing levels which have essentially the same effect on the response. To encourage effect fusion, we propose a

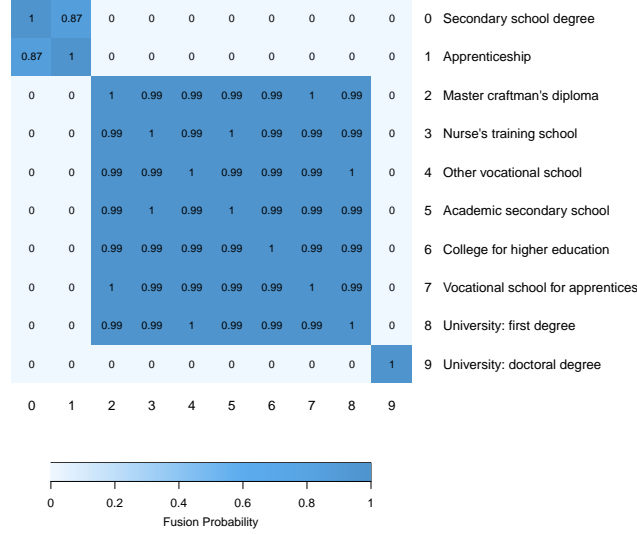


Figure 4: Covariate **education**: estimated pairwise fusion probabilities of level effects. Category 0 indicates the baseline level. Darker colours indicate higher fusion probabilities.

Normal prior distribution that allows for almost perfect as well as almost zero partial dependence of level effects. Alternatively, this prior can be derived as a spike and slab prior on all level effect contrasts associated with one covariate and taking the linear restrictions among them into account. As an advantage the construction of the prior easily allows to incorporate prior information on which pairs of levels should not be fused directly. This property is of particular interest for ordinal covariates where typically fusion would be restricted to subsequent levels.

Posterior inference using MCMC methods is straightforward. Model selection can be based on the estimated posterior means of the pairwise fusion probabilities. To avoid selection of logically inconsistent models we suggest to fuse effects when posterior evidence is clear. Simulation results show that the proposed method automatically excludes irrelevant predictors totally and outperforms competing methods in terms of correct model selection, coefficient estimation as well as prediction.

Bayesian effect fusion is not restricted to categorical predictors in linear regression models but can be applied also in more general regression models e.g. generalised linear models. Only little adaption is required for posterior simulation in a Bayesian regression type model, where posterior inference using MCMC methods is feasible for a Normal prior on the regression effects.

A certain drawback of the method is that to construct the prior covariance matrix all pairwise effect differences have to be assessed in each MCMC sampling step and hence the computational effort can be prohibitive for nominal covariates with a very high number of levels.

Acknowledgement: This work was financially supported by the Austrian Science Fund (FWF) via the research project number P25850 ‘Sparse Bayesian modelling for categorical predictors’.

References

- Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* 65, 169–177.
- Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics* 1, 17–36.
- Fahrmeir, L., T. Kneib, and S. Konrath (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing* 20, 203–219.
- George, E. and R. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 7, 339–373.
- Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with application to international classification of functioning score sets. *Journal of Royal Statistical Society, Series C*, 377 – 395.
- Gertheiss, J. and G. Tutz (2009). Penalized regression with ordinal predictors. *International Statistical Review*, 345 –365.
- Gertheiss, J. and G. Tutz (2010). Sparse modelling of categorical explanatory variables. *The Annals of Applied Statistics* 4, 2150 – 2180.
- Griffin, J. and P. J. Brown (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 5, 171–188.
- Huang, A., S. Xu, and X. Cai (2015). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. *Heredity* 114(1), 107–115.

- Ishwaran, H., L. F. James, and J. Sun (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 1316–1332.
- Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized regression, standard errors, and Bayesian lasso. *Bayesian Analysis* 5(2), 369–412.
- Liu, F., S. Chakraborty, F. Li, Y. Liu, and A. C. Lozano (2014). Bayesian regularization via graph Laplacian. *Bayesian Analysis* 9(2), 449–474.
- Mitchell, T. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83, 1023 – 1032.
- Park, T. and G. Casella (2008). The Bayesian lasso. *Journal of the American Statistical Association* 103(482), 681–686.
- Raman, S., T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth (2009). The bayesian group-lasso for analyzing contingency tables. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML 2009, Montreal.
- Rue, H. and L. Held (2005). *Gaussian Markov Random Fields. Theory and Applications*. Chapman and Hall/CRC.
- Scheipl, F., L. Fahrmeir, and T. Kneib (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association* 107(500), 1518–1532.
- Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* 22:2, 231–245.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B* 58(1), 267–288.
- Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Society, Series B* 67(1), 91–108.
- Tutz, G. and M. Berger (2014). Tree-structured modelling of categorical predictors in regression. Technical report, Cornell University Library. arXiv:1504.04700.
- Tutz, G. and J. Gertheiss (2016). Regularized regression for categorical data. *Statistical Modelling* 16(3), 161–200.

- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society, Series B* 68, 49–67.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, Series B* 67, 301–320.

A Appendix

A.1 Properties of the effect fusion prior

\mathbf{Q} is symmetric by definition. To show positive definiteness of \mathbf{Q} we consider the quadratic form where \mathbf{x} is a nonzero vector of dimension c .

Obviously, we have

$$\begin{aligned}
\mathbf{x}'\mathbf{Q}\mathbf{x} &= \sum_{k=1}^c \sum_{j=1}^c x_k x_j q_{kj} = \\
&= \sum_{k=1}^c x_k^2 \left(\sum_{j \neq k} \kappa_{kj} \right) - \sum_{k=1}^c \sum_{\substack{j=1 \\ j \neq k}}^c x_k x_j \kappa_{kj} = \\
&= \sum_{k=1}^c x_k^2 \kappa_{k0} + \sum_{k=1}^c \sum_{\substack{j=1 \\ j \neq k}}^c x_k^2 \kappa_{kj} + \sum_{k=1}^c \sum_{\substack{j=1 \\ j \neq k}}^c x_k x_j \kappa_{kj} = \\
&= \sum_{k=1}^c x_k^2 \kappa_{k0} + \sum_{k=1}^c \sum_{0 < j < k}^c (x_k - x_j)^2 \kappa_{kj} > 0
\end{aligned}$$

if $\mathbf{x} \neq \mathbf{0}$, where $\mathbf{0}$ is a vector of zeros.

A.2 Spike and slab priors on effect differences

We show that the prior on the regression effects $\boldsymbol{\beta}$ given in (3) with prior precision matrix (5) corresponds to specifying independent spike and slab priors on all effect differences and then correcting for the linear restrictions on these effect differences. Spike and slab priors on effect differences $\theta_{kj} = \beta_k - \beta_j$ are specified as

$$\begin{aligned}
\theta_{kj} | \delta_{kj}, \tau^2 &\sim \delta_{kj} \mathcal{N}(0, \tau^2 \gamma) + (1 - \delta_{kj}) \mathcal{N}\left(0, \frac{1}{r} \tau^2 \gamma\right) \\
\tau^2 &\sim \mathcal{G}^{-1}(g_0, G_0).
\end{aligned}$$

Conditional on the hyperparameters, the prior on the effect difference θ_{kj} can be written more compactly as

$$\theta_{kj} \sim \mathcal{N}\left(0, \frac{\tau^2}{\kappa_{kj}} \gamma\right).$$

We subsume all effect differences θ_{kj} in the $d \times 1$ vector $\boldsymbol{\theta}$ so that the first c elements correspond to $\boldsymbol{\beta}$ and partition $\boldsymbol{\theta}$ accordingly in $\boldsymbol{\theta} = (\boldsymbol{\beta} = \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. We write the linear restrictions

$$\theta_{kj} = \beta_k - \beta_j \iff -\theta_{j0} + \beta_{k0} - \theta_{kj} = 0$$

in matrix form as $\mathbf{L}\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 = \mathbf{0}$.

The distribution of a Normal vector $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ under the linear restriction $\mathbf{R}\boldsymbol{\theta} = \mathbf{0}$ is again Normal with moments

$$\begin{aligned} \mathbb{E}(\boldsymbol{\theta} | \mathbf{R}\boldsymbol{\theta} = \mathbf{0}) &= \mathbf{0} \\ \text{Cov}(\boldsymbol{\theta} | \mathbf{R}\boldsymbol{\theta} = \mathbf{0}) &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{R}'(\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}')^{-1}\mathbf{R}\boldsymbol{\Sigma} \end{aligned}$$

see Rue and Held (2005), p. 37.

To determine the covariance matrix of $\boldsymbol{\theta}$ conditioning on $\mathbf{R} = (\mathbf{L} \quad -\mathbf{I})$, where \mathbf{I} is the identity matrix of dimension $d - c$, we partition also $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \tau^2\gamma\boldsymbol{\Delta} = \tau^2\gamma \begin{pmatrix} \boldsymbol{\Delta}_1 & \\ & \boldsymbol{\Delta}_2 \end{pmatrix}$$

which yields

$$(\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}')^{-1} = \frac{1}{\tau^2\gamma}(\mathbf{L}\boldsymbol{\Delta}_1\mathbf{L}' + \boldsymbol{\Delta}_2)^{-1}.$$

The covariance matrix of $\boldsymbol{\theta}$ under the linear restrictions is then

$$\begin{aligned} \text{Cov}(\boldsymbol{\theta} | \boldsymbol{\theta}_2 - \mathbf{L}\boldsymbol{\theta}_1 = \mathbf{0}) &= \tau^2\gamma \left(\boldsymbol{\Delta} - (\boldsymbol{\Delta}_1\mathbf{L}' \quad -\boldsymbol{\Delta}_2) \mathbf{W} \begin{pmatrix} \mathbf{L}\boldsymbol{\Delta}_1 \\ -\boldsymbol{\Delta}_2 \end{pmatrix} \right) = \\ &= \tau^2\gamma \left[\begin{pmatrix} \boldsymbol{\Delta}_1 & \\ & \boldsymbol{\Delta}_2 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Delta}_1\mathbf{L}'\mathbf{W}\mathbf{L}\boldsymbol{\Delta}_1 & -\boldsymbol{\Delta}_2\mathbf{W}\mathbf{L}\boldsymbol{\Delta}_1 \\ -\boldsymbol{\Delta}_1\mathbf{L}'\mathbf{W}\boldsymbol{\Delta}_2 & \boldsymbol{\Delta}_2\mathbf{W}\boldsymbol{\Delta}_2 \end{pmatrix} \right], \end{aligned}$$

where $\mathbf{W} = (\mathbf{L}\boldsymbol{\Delta}_1\mathbf{L}' + \boldsymbol{\Delta}_2)^{-1}$. As \mathbf{B}_0 of $\boldsymbol{\beta} = \boldsymbol{\theta}_1$ is the upper left $c \times c$ matrix of $\text{Cov}(\boldsymbol{\theta} | \boldsymbol{\theta}_2 = \mathbf{L}\boldsymbol{\theta}_1)$, we obtain

$$\mathbf{B}_0 = \text{Cov}(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2 = \mathbf{L}\boldsymbol{\theta}_1) = \tau^2\gamma (\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_1\mathbf{L}'\mathbf{W}\mathbf{L}\boldsymbol{\Delta}_1).$$

The structure matrix \mathbf{Q} therefore is given as

$$\mathbf{Q} = (\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_1\mathbf{L}'\mathbf{W}\mathbf{L}\boldsymbol{\Delta}_1)^{-1}$$

and can be simplified to

$$\mathbf{Q} = \boldsymbol{\Delta}_1^{-1} + \mathbf{L}'\boldsymbol{\Delta}_2^{-1}\mathbf{L}$$

using the Woodbury formula.

Thus, the off-diagonal elements of \mathbf{Q} are

$$q_{kj} = \ell'_k \Delta_2^{-1} \ell_j,$$

where ℓ_k denotes the k -th column of \mathbf{L} . For each pair of columns ℓ_k and ℓ_j , there is exactly one row where both vectors have a non-zero element, which takes the value 1 for one and -1 for the other vector and therefore we have

$$q_{kj} = -\kappa_{kj}.$$

Finally, the diagonal elements of \mathbf{Q} are given as

$$q_{ii} = \kappa_{i0} + \ell'_i \Delta_2^{-1} \ell_i = \kappa_{i0} + \sum_{\substack{j=1 \\ j \neq k}}^c \kappa_{kj}.$$

B Details on the simulation study

B.1 Alternative methods

Table 6 lists the methods to which we compare Bayesian effect fusion in the simulation study, together with the name of the corresponding R packages and the references given in the package manuals. The code of the Graph Laplacian approach in Liu et al. (2014) was provided directly from the authors.

Table 6: Details for methods used in simulation study

Method	R package	References
Penalty	gvcn.cat	Gertheiss and Tutz (2010)
BLasso	monomvn	Park and Casella (2008)
BEN	EBglmNet	Huang et al. (2015)
GLasso	grpreg	Yuan and Lin (2006)
GLap	-	Liu et al. (2014)

B.2 Model selection measures

As measures for correct model selection, we use true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV) and negative predictive

value (NPV). Generally, they are defines as follows:

$$\text{TPR} = \text{TP}/P = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{TNR} = \text{TN}/N = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{NPV} = \text{TN}/(\text{TN} + \text{FN}).$$

In our setting of level effect fusion, TP (*true positive*) is the number of correctly detected non-zero effect differences, (TN) *true negative* the number of correctly detected zero difference, FN (*false negative*) the number of zero effect difference clasisified as npn-zero and (FP) *false positive* the number of zero effect differences classified as non-zero.